Guest Editorial

PERCEPTION

Why We Need to Do Fewer Statistical Tests

Perception
2016, Vol. 45(5) 489–491

DOI: 10.1177/0301006616637434
pec.sagepub.com
SAGE

A recent collaboration, in which a large number of psychological studies were carefully repeated, found that a majority of the findings could not be replicated (Open Science Collaboration, 2015). Is there any reason to believe that articles in *Perception* are more reliable? If not, is there anything we can do to increase the reliability? In order to answer this question, we need to consider why so many findings could not be replicated. Presumably, this is because the findings were not really true effects in the first place, despite being significant. They were *false positives*. But why are there so many false positives? Cases of scientists fabricating data to prove their point are disturbing, but not common, so there must be some other reason. Questionable practices such as testing more subjects when an effect is close to significance can certainly make reported statistics less reliable (Simmons, Nelson, & Simonsohn, 2011), but this is unlikely to be responsible for the fact that over half the effects are not reproducible.

Could the abundance of false positives have anything to do with the fact that significant effects are overrepresented in the scientific literature (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014)? The overrepresentation of significant effects is presumably mainly the result of various biases in reporting (Ioannidis et al., 2014). Authors may not bother trying to publish data if the effect that they were looking for was not significant. Moreover, reviewers are often more critical when evaluating papers in which the null hypothesis was not rejected. This is not completely unjustified, because claiming that there is an effect when it is significant is rather straightforward, whereas claiming that there is no effect when it is not significant requires an estimate of the expected effect size and of the variability in the measure of interest, in order to guarantee that the effect would have been significant if it were present. Moreover, there has to be some justification for expecting an effect in the first place. It is perhaps not directly obvious why not publishing some papers would make the others less reproducible, but I will explain why I believe that the underlying emphasis on significant effects does have severe consequences when combined with the tendency to use statistics to explore one's own data.

Our standard statistical tests were developed to test hypotheses. As a community, we accept a 5% chance of false positives. The reason for being so lenient is that being very strict would make us often conclude incorrectly that there is no effect when actually there *is* an effect. Hopefully, 5% is a reasonable value in terms of matching the likelihoods of incorrectly concluding that there is an effect when there is none (a Type 1 error), and of incorrectly concluding that there is no effect when actually there is an effect (a Type 2 error). Accepting a 5% chance of making a Type 1 error might make us expect that only 5% of attempts to replicate significant effects will fail. However, it is not that simple. One way to see this is by considering a field of research in which people only test things that we know to have no effect. If there is no true effect and the measures all have normally distributed variability, then about 5% of the tests will be significant. If we select the effects that appear to be significant and try to replicate them, obviously only about 5% of them will replicate.

Thus, if we select the significant effects from many tests of things that certainly have no true effect, we will fail to replicate about 95% of them. The Open Science Collaboration only tried to replicate significant or almost-significant effects.

It is safe to assume that in our field people do not *only* test things that have no true effect, and true effects obviously have a much higher likelihood of being replicable. However, it is just as safe to assume that we cannot only test effects that are true effects, because if we know which effects are true effects we do not need to do the research. Testing things that have no true effect inevitably gives rise to false positives, and the higher the proportion of false positives, the lower the probability of a randomly chosen significant effect being replicable. Thus, the more we test for effects that are not true effects, the less reproducible "significant effects" will be (for a more thorough explanation see Ioannidis, 2005).

What can be done about this? It is certainly not a good idea to discourage people from testing well-motivated hypotheses. However, asking for fewer rather than more statistical tests might help. Imagine that we are interested in whether metal objects look heavier than wooden ones. We decide to test this by presenting subjects with a 1-cm diameter metal sphere next to a wooden sphere of one of several diameters and asking them to indicate which looks heavier (a standard two alternative forced choice task). Just in case objects on the right look heavier than ones on the left, we randomly pick the side on which the metal sphere is presented. We then use standard techniques to determine the size of the wooden sphere that matches the metal sphere in apparent weight (for instance, by fitting a psychometric curve to the fractions of times the wooden sphere was judged to be heavier for various sizes of the wooden sphere and determining the size for which the fraction would be exactly half). If we determine this for several people, we could use a *t*-test to see whether the material has a systematic influence on the judged weight (in which case the size will differ systematically from that of the metal sphere). Presumably it does. So presumably, we would have a significant effect, and it would be no problem to replicate this effect. Since the metal sphere is sometimes on the left and sometimes on the right, we might be lured into testing whether the influence of the material depends on the side on which the metal sphere was presented. We might also notice during the experiment that several participants who wear spectacles have particularly large effects and therefore decide to check whether the effect depends on whether the participant wears spectacles. Moreover, we may have reported that we had both male and female participants, and a reviewer may have asked us to test whether there were any gender differences. By testing whether these factors and their interactions influence the effect of material on the judged weight, we will have added a large number of tests of effects that presumably are not present. A full three-way ANOVA (with factors side, spectacles and gender) will test seven additional options (three main effects, three two-way interactions, and a three-way interaction).

Adding this kind of ANOVA is often considered to be a good thing because you are getting more out of your data. The problem with carrying out such ANOVAs is that it reduces the overall reliability of significant results. If we want to increase the reproducibility of reported effects, the simplest thing to do is to stop testing for effects that are not part of the hypothesis under study. This does not mean that one should not look critically at the data to see whether there are any relationships that might have influenced the findings. However, if one finds something that looks interesting, one should not use statistics that were devised to test hypotheses to decide whether the effect should be taken seriously. For every question that is being asked, there will usually be a critical statistical measure that needs to be reported. In many studies, various possible confounding factors need to be considered. Providing data about them is fine, but there is no need to run statistical tests on all of them. If it seems from the data that people wearing spectacles judge metal objects to be heavier in relation to

wooden ones than do people who do not wear spectacles, one might want to look into this, but conducting the additional ANOVA mentioned above is not the way to test this unless you started off with the hypothesis that spectacles would have this effect.

The fact that so many reported effects cannot be replicated means that it is worthwhile trying to replicate important findings before relying on them. Consequently, if you fail to replicate an important effect you should not be expected to explain why you got a different result, as is now often the case, but it should be acceptable to propose that the effect might have been a statistical coincidence. Publishing such failures to replicate will help prevent false positives from guiding our theories and further research. If a replication does give the same result, it should also get published, because replications indicate that the particular finding should guide our theories and further research. However, rather than trying to publish many replications, it might be better to strive for fewer false positives to start with. The easiest way to reduce the number of false positives is by discouraging exploratory statistical testing. If an experiment is testing a hypothesis, there should seldom be a need for more than one comparison from one statistical test in order to determine whether to reject the hypothesis. Drawing any conclusions from additional comparisons or tests should therefore be considered suspicious. I would therefore like to encourage authors to use fewer statistical tests, and reviewers and editors to ask for fewer, rather than more, statistical tests. It is important to stop building theories on spurious significant findings, so we, as scientists, can spend our time figuring out how to interpret the findings rather than arguing about the data.

Eli Brenner
Department of Human Movement Sciences, VU University, Amsterdam, The Netherlands.
Email: e.brenner@fbw.vu.nl

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Editorial Note

We welcome comments on our editorials, which may be published in a later issue subject to editorial review. Please send comments to Gillian Porter at gillian.porter@bristol.co.uk.

References

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*, e124.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Science*, *18*, 235–241.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.